Proseminar „Verarbeitung geographischer Daten (Quant II)“

Sommersemester 2001

Daniel Braunschweiger

Achim Schmidt

Tobias Spaltenberger

Diskriminanzanalyse

– am Fallbeispiel von 23 Klimastationen in Tunesien –

Inhaltsverzeichnis

1 Einführung

1.1 Was ist die Diskriminanzanalyse?

Die Diskriminanzanalyse ist ein multivariates Verfahren zur Analyse von Gruppen- bzw. Klassenunterschieden. Mit dieser Methode ist es möglich, zwei oder mehrere Gruppen unter Berücksichtigung von mehreren Variablen zu untersuchen und zu ermitteln, wie sich diese Gruppen unterscheiden.

Im Unterschied zur Clusteranalyse ist die Diskriminanzanalyse kein exploratives, sondern ein konfirmatorisches Verfahren. Durch die Diskriminanzanalyse werden keine Gruppen gebildet, sondern man geht von einer vorhandenen Gruppierung aus und überprüft die Qualität dieser Gruppierung.

Durch die Diskriminanzanalyse lässt sich analysieren,

- ob die vorliegende, möglicherweise durch Clusteranalyse ermittelte Gruppierung optimal ist oder ob sie verbessert werden kann;

- welche Variablen für die Gruppenbildung besonders geeignet sind bzw. auf welche Variablen sich die Gruppenunterschiede hauptsächlich zurückführen lassen;
- in welche Gruppe ein neues Objekt aufgrund seiner Merkmalsausprägungen einsortiert werden kann.

1.2 Verwendung der Diskriminanzanalyse in der Geographie

Für die Geographie ist die Diskriminanzanalyse vor allem bei der Lösung der folgenden Aufgabenstellungen einsetzbar (BAHRENBERG, 1992, S. 318):

- Trennung von Raumeinheiten nach verschiedenen Merkmalen zur Raumtypisierung;
- Überprüfung und u.U. Verbesserung einer vorgegeben Raumgliederung, die z.B. per Clusteranalyse ermittelt wurde;
- Zuordnung von nicht klassifizierten Raumeinheiten zu vorgegeben Raumtypen;
- Analyse der Unterschiede zwischen verschieden Raumtypen.

1.3 Voraussetzung für die Diskriminanzanalyse

- metrisch skalierte Merkmalsvariablen
- optimalerweise eine Normalverteilung der vorliegenden Daten
- kein Element der Stichprobe darf gleichzeitig mehreren Gruppen zugeordnet sein
- der Stichprobenumfang soll mindestens doppelt so hoch sein wie die Anzahl der Merkmalsvariablen
- die Anzahl der Merkmalsvariablen sollte größer sein als die Anzahl der Gruppen

2 Verfahren

2.1 Definition der Gruppen und Variablen

Zuerst muß definiert werden, welche Gruppen nun anhand welcher Variablen überprüft werden sollen.Wir übernehmen das Beispiel der Clusteranalyse, mit deren Hilfe 23 Klimastationen Tunesiens in 2 Gruppen A u. B eingeteilt wurden. Diese Gruppierung soll nun anhand der 2 Variablen „Niederschlag“ und „Höhe über NN“ überprüft werden.

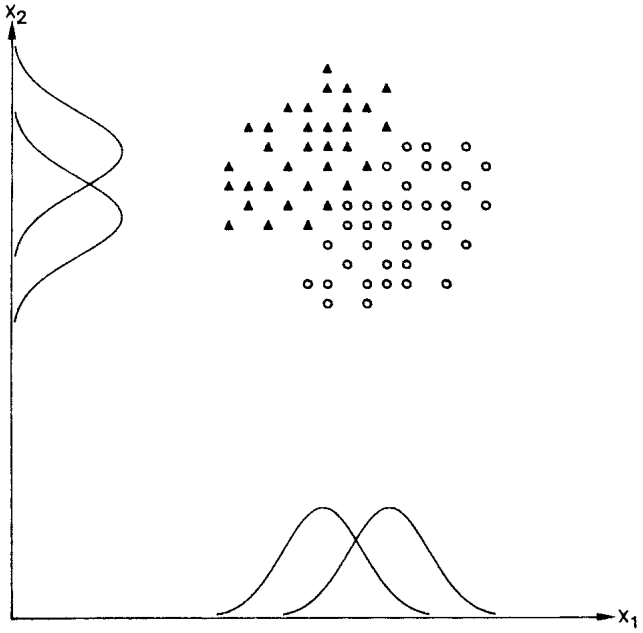

Abb. 1: Trennung durch die Ausgangsvariablen x_1 und x_2. (Quelle: BAHRENBERG (1992): 319)

2.2 Formulierung der Diskriminanzfunktion (Trennfunktion)

Gesucht wird nun eine Funktion, die die Gruppen optimal trennt. In **Abb. 1** wurden die Häufigkeitsverteilungen der beiden Gruppen A (Dreiecke) und B (Kreise) jeweils auf die x_1- bzw. x_2-Achse projiziert, erkennbar sind rel. große Überschneidungsbereiche, d.h.: die Werte im Überschneidungsbereich können weder Gruppe A noch Gruppe B definitiv zugeordnet werden. Die Variablen (die Achsen) x_1 u. x_2 sind also als Trennfunktionen nicht geeignet. Auch in **Abb. 2** sind für die Funktionen Y* und Y** große Überschneidungsbereiche erkennbar, d.h.: unsaubere Trennung der beiden Gruppen. Funktion Y dagegen weist

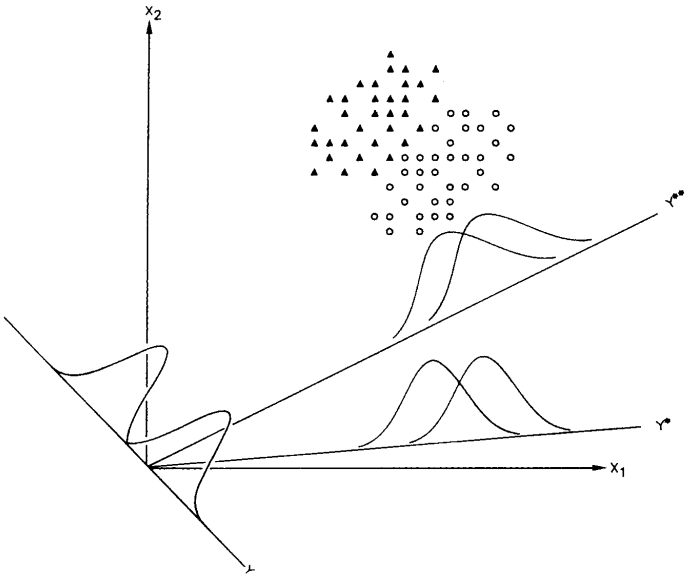

Abb. 2: Trennung durch verschiedene Diskriminanzachsen (Quelle: BAHRENBERG (1992): 321)

keinerlei Überschneidungsbereich auf, eine Trenngerade kann eingezeichnet werden, die die Gruppen optimal trennt (**Abb. 3**), Y ist also die gesuchte Diskriminanzfunktion. Die Funktion läßt sich als Linearfunktion der beiden Merkmalsvariablen x_1 u. x_2 beschreiben mit der Gleichung

$$Y = v_1 x_1 + v_2 x_2 .$$

Die Lage der Diskriminanzachse im Raum, also ihre Steigung, ist durch das Verhältnis

$$x_2 = \frac{v_2}{v_1} x_1$$ bestimmt.

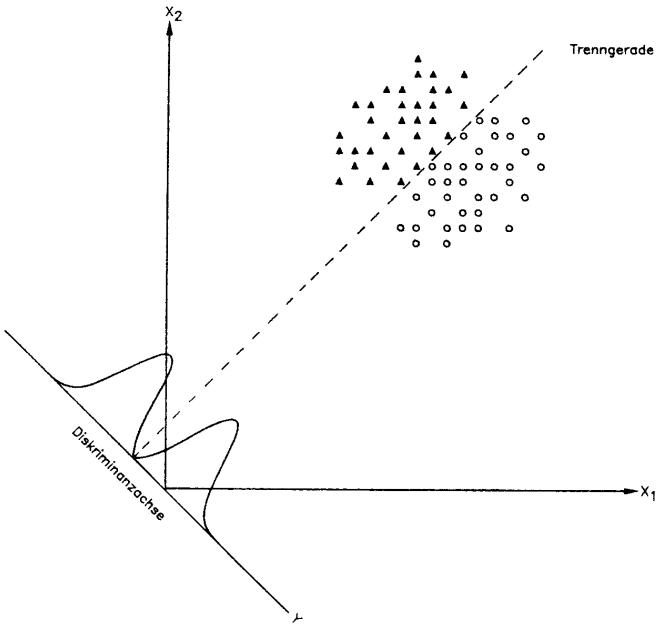

Abb. 3: Trennung durch die Diskriminanzachse (Quelle: BAHRENBERG (1992): 321)

Vereinfachend wird angenommen, die Gerade verlaufe durch den Ursprung.
→ Die Diskriminanzkoeffizienten v_1 u. v_2 sind nun aufgrund der Daten der Merkmalsausprägungen zu schätzen.

2.3 Schätzung der Diskriminanzfunktion

Wie oben ausgeführt, bedingt gute Trennung kleine Überschneidungsbereiche der Häufigkeitsverteilungen auf der Diskriminanzachse. Dieses kann mit Hilfe zweier Kriterien erreicht werden:

1. Maximierung der Streuung *zwischen* den Gruppen(max. der Abstände der Gruppenmittelpunkte auf der Diskriminanzachse):

$$d = \left| \overline{y_A} - \overline{y_B} \right| \quad \text{bzw.} \quad d^2 = \left(\overline{y_A} - \overline{y_B} \right)^2$$

2. Minimierung der Streuung *innerhalb* der Gruppen:

$$s^2 = \sum_{j=1}^{n_A} \left(y_{Aj} - \overline{y_A} \right)^2 + \sum_{j=1}^{n_B} \left(y_{Bj} - \overline{y_B} \right)$$

Hieraus leitet sich das Diskriminanzkriterium wie folgt ab: $\Gamma = \frac{d^2}{s^2} \rightarrow \max$

Durch Maximierung dieses Kriteriums, d.h. durch Maximierung des Abstandes *zwischen* den Gruppen und Minimierung der Abstände *innerhalb* der Gruppen sind die Bedingungen für einen möglichst kleinen Überschneidungsbereich

erfüllt. Nach partieller Differentiation von Γ nach v_1 und v_2 erhält man die Bestimmungsgleichungen, die zur Berechnung der Diskriminanzfunktion benötigt werden. Durch Einsetzen der Werte für x_1 u. x_2 in diese Gleichungen können die Werte für v_1 u. v_2 berechnet werden und erhält somit die gesuchte Geradengleichung.

2.4 Prüfung der Diskriminanzfunktion

Bei der Prüfung der Diskriminanzfunktion wird deren Güte bzw. Trennkraft hinsichtlich der verschiedenen Gruppen untersucht. Dabei wird die Unterschiedlichkeit der Gruppen gemessen. Es gibt zwei Möglichkeiten, eine Diskriminanzfunktion zu messen:

a) Vergleich mit kritischem Diskriminanzwert

Für jedes Element j lässt sich ein Wert auf der Diskriminanzachse berechnen:

$$y_j = v_1 x_{1j} + v_2 x_{2j}$$

Diese errechneten Diskriminanzwerte werden nun mit einem sogenannten kritischen Diskriminanzwert verglichen. Dieser Wert ist bei zwei gleichgroßen Clustern nichts anderes als das arithmetische Mittel der Diskriminanzwerte y_j:

$$y_t = \frac{\overline{y_A} + \overline{y_B}}{2}$$

Kritischer Diskriminanzwert

arithmetisches Mittel von Cluster B

arithmetisches Mittel von Cluster A

Bei zwei unterschiedlich großen Clustern muss man das gewichtete arithmetische Mittel berechnen:

$$y_t = \frac{n_A \overline{y_A} + n_B \overline{y_B}}{n_A + n_B}$$

Anzahl der Elemente von Cluster A Anzahl der Elemente von Cluster B

Nachdem man nun den kritischen Diskriminanzwert errechnet hat, kann man die einzelnen Diskriminanzwerte in die Cluster einteilen. Diese Einteilung hängt davon ab, ob $y_j < y_t$ oder $y_j > y_t$ ist.
Eine quantitative Aussage über die Trennkraft einer Diskriminanzfunktion läßt sich anhand einer sogenannten Klassifikationsmatrix treffen:

		Gruppenzugehörigkeit nach der Diskriminanzanalyse			
		Cluster A		*Cluster B*	
Vorgegebene Gruppenzugehörigkeit	*Cluster A*	**7** (87,5%)	+	1 (12,5%)	= 8
		+		+	+
	Cluster B	0	+	**8** (100%)	= 8
		= 7	+	= 9	=16

Abb. 4: Klassifikationsmatrix einen Fallbeispieles (Quelle : BAHRENBERG (1999) : 328)

Auf der sogenannten Hauptdiagonalen, im obigen Fall die hervorgehobenen Werte, sind die korrekt klassifizierten Elemente dargestellt. In diesem Fallbeispiel war also nur ein einziges Element falsch klassifiziert. Mit den richtig klassifizierten Elementen läßt sich jetzt die „Trefferquote“ errechnen. Diese beträgt hier 93,75%. Diese Trefferquote sagt uns etwas über die tatsächliche Trennkraft der Diskriminanzfunktion aus.
Wichtig: Nur wenn die Trefferquote größer als 50% ist, d.h. größer als nach dem Zufallsprinzip zu erwarten wäre, ist eine Diskriminanzfunktion von Nutzen.

b) Betrachtung des Diskriminanzkriteriums

In diesem Fall bildet der Eigenwert (Maximalwert des Diskriminanzkriteriums) ein Maß für die Güte, bzw. Trennkraft einer Diskriminanzfunktion:

$$\gamma = \frac{SS_b}{SS_w}$$

Eigenwert

erklärte Streuung

nicht erklärte Streuung

Da der Eigenwert nicht auf Werte zwischen null und eins normiert ist, muss man sich andersweitig aushelfen. Es gibt zwei Möglichkeiten, diesem Problem aus dem Weg zu gehen. Die erste Möglichkeit ist die Berechnung des kanonischen Korrelationskoeffizienten:

$$c = \sqrt{\frac{\gamma}{1+\gamma}} = \sqrt{\frac{SS_b}{SS_b + SS_w}}$$

Erklärte Streuung

Gesamtstreuung

In der Diskriminanzanalyse wird dieser Wert als Gütemaß verwendet. Das Besondere am kanonischen Korrelationskoeffizienten ist, dass dieser Wert im Zwei-Gruppen-Fall dem Bestimmtheitsmaß einer Regressionsanalyse entspricht.

Wichtig: Je größer der kanonische Korrelationskoeffizient ist, desto höher ist die Trennkraft der Diskriminanzfunktion!

Eine zweite Möglichkeit, dass man normierte Werte zwischen null und eins erhält, ist die Berechnung des sogenannten Wilks' Lambda :

$$\Delta = \frac{1}{1+\gamma} = \frac{SS_w}{SS_b + SS_w}$$

nicht erklärte Streuung

Gesamtstreuung

Im Gegensatz zum kanonischen Korrelationskoeffizienten handelt es sich hierbei um ein „inverses“ Gütemaß, d.h. je kleinere Werte Wilks’ Lambda annimmt, desto größer ist die Trennkraft einer Diskriminanzfunktion.
Das Wilks’ Lambda ist laut BACKHAUS (2000) die gebräuchlichste Methode zur Prüfung einer Diskriminanzfunktion.
Der Vorteil liegt darin, dass auch Signifikanztests über die Unterschiedlichkeit der Gruppen möglich sind. Dies wird im folgenden aber nicht weiter ausgeführt.
Fazit: Erst wenn man eine geschätzte Diskriminanzfunktion hinsichtlich ihrer Trennkraft untersucht hat, kann man mit ihr weiter verfahren. Stellt man aber fest, dass ihre Trennkraft nicht sonderlich hoch ist, macht es keinen Sinn mit ihr z. B. neue Objekte zu klassifizieren (s.u.) oder, wie im folgenden dargestellt die Merkmalsvariablen zu überprüfen.

2.5 Prüfung der Merkmalsvariablen

Hiebei will man überprüfen, welche der beiden Merkmalsvariablen einen größeren Einfluß auf die Einteilung der Cluster ausübt. Die Basis für die Beurteilung der Trennkraft der Variablen bilden die Diskriminanzkoeffizienten v, da sie den Einfluß einer Merkmalsvariable auf den Diskriminanzwert repräsentiert. Es gibt allerdings jetzt ein Problem: Die Koeffizienten werden von der Maßeinheit der Ausgangsvariablen beeinflußt, d.h. sie reagieren auf Skalierungen. Darum werden die Diskriminanzkoeffizienten mit den Standardabweichungen der entsprechenden Variablen multipliziert.
Im folgenden wird die Formel für den standardisierten Diskriminanzkoeffizienten einer beliebigen „Merkmalsvariablen 1“, beispielsweise Niederschlag in mm, aufgestellt. Der Rechenvorgang für die dazugehörige „Merkmalsvariablen 2“ läuft entsprechend gleich ab.

$$c_1 = v_1 s_{x_1}$$

Standardisierter Diskriminanzkoeffizient

Diskriminanz-koeffizient

Standardabweichung der Merkmalsvariablen

Durch diesen Rechenschritt sind die Diskriminanzkoeffizienten nun standardisiert, und man kann nun mit der Prüfung der Merkmalsvariablen fortfahren.
Man hat nun für beide Cluster den standardisierten Korrelationskoeffizienten vorliegen. Nun vergleicht man beide Werte miteinander.
Wichtig: Diejenige Merkmalsvariable, die einen größeren standardisierten Diskriminanzkoeffizienten hervorruft hat einen größeren Einfluß auf die Clustertrennung.
Auf das Maß für den Einfluß auf die Trennung kann man schließen, wenn man den Unterschied der beiden Werte betrachtet: Je größer dieser zwischen den beiden Werten ist, desto ausgeprägter ist die Einflußnahme der Merkmalsvariablen mit dem größeren standardisierten Diskriminanz-koeffizienten.

2.6 Klassifizierung von Objekten

Nachdem man nun die Diskriminanzfunktion, sowie auch die Merkmalsvariablen, überprüft hat, hat man nun die Möglichkeit mit Hilfe dieser neue Objekte zu klassifizieren, d. h. in eine der beiden Clustern einzuordnen.
Dies erreicht man in dem, wie bei der Prüfung der Diskriminanzfunktion dargestellt, die Merkmalsvariablen des neuen Objektes in die Diskriminanzfunktion eingesetzt und der errechnete Wert mit dem schon bekannten kristischen Diskriminanzwert verglichen wird. Je nachdem ob der neue Diskriminanzwert größer oder kleiner als der kritische Diskriminanzwert ist, wird dieses neue Objekt in einen der beiden Cluster eingeteilt.

2.7 Mehr-Gruppen – Mehr-Variablen-Fall

Der in Praxis am häufigsten auftretende Fall ist nicht der oben beschriebene Zwei-Gruppen–Zwei-Variablen-Fall, häufiger ist der Mehr-Gruppen–Mehr-Variablen-Fall.
Bei diesem Fall reicht eine Diskriminanzfunkion nicht mehr aus, um die Gruppen zufriedenstellend zu trennen. Nach der Ermittlung der ersten Diskriminanzachse sind in der Regel die Überlappungsbereiche so groß, daß

weitere Achsen bestimmt werden können, bei *G* Gruppen maximal *G - 1* Diskriminanzachsen. Da die Anzahl der Diskriminanzfunktionen nicht höher sein sollte als die Anzahl der Merkmalsvariablen *I*, ist die maximale Anzahl der Diskriminanzfunktionen auf *K = Min (G – 1, I)* festgelegt. Im Normalfall reichen jedoch zwei Diskriminanzfunktionen aus.

Zur Beurteilung der Bedeutung der einzelnen Diskriminanzfunktion für das Trennverfahren wird der Eigenwertanteil jedes Eigenwertes EA_k, der erklärte Varianzanteil verwendet:

$$EA_k = \frac{\gamma_k}{\gamma_1 + \gamma_2 + \gamma_3 + ... + \gamma_k} \quad \text{mit} \quad k = 1, 2, 3, ..., K$$

Er gibt die durch die *k*-te Diskriminanzfunktion erklärte Streuung als Anteil der Gesamtstreuung an.

Zur Klassifizierung von Objekten beim Mehr-Gruppen–Mehr-Variablen-Fall gibt es mehrere Konzepte, unter anderem das Distanzkonzept und das Wahrscheinlichkeitskonzept. Beim Distanzkonzept wir ein Element derjenigen Gruppe zugeordnet, zu deren Gruppenmittelpunkt (Zentroid) es den geringsten Abstand hat. Das Wahrscheinlichkeitskonzept ist eine Weiterentwicklung des Distanzkonzepts. Dabei wird ein Element mit dem Diskriminanzwert y_j derjenigen Gruppe *g* zugeordnet, bei der die Wahrscheinlichkeit $p\ (gly_j)$ maximal ist. Die Klassifizierungswahrscheinlichkeit werden nach dem Bayes-Theorem bestimmt.

3 Fallbeispiel Tunesien

Als Fallbeispiel sind 23 Klimastationen Tunesiens mit ihrer Höhe über NN und ihrem durchschnittlichen Jahresniederschlag ausgewählt worden. Sie wurden mittels Clusteranalyse in 2 Gruppen (A und B, siehe auch **Abb. 4**) eingeteilt:

STATION	Höhe über NN (m)	Niederschlag (mm)	Gruppe
Medjaz El-Bab	54	408	A
Kelibia	82	434	A
Soliman	12	458	A
Grombalia	50	475	A
Zaghouan	195	496	A
Le Kef	665	509	A
Souk El-Arba	143	449	A
Maktar	934	490	A
Thala	1.020	473	A
Bizerba	02	625	A
Beja	234	626	A
Teboursouk	410	523	A
Tabarka	12	1.029	A
Ain-Draham	739	1.534	A
Sousse	06	327	B
Kairouan	68	286	B
El Djem	112	267	B
Gafsa	314	152	B
Sfax	21	197	B
Gabes	02	175	B
Kebili	56	89	B
Tozeur	46	89	B
Metlaui	234	137	B

Die nun folgenden Daten wurden mit SPSS 10 berechnet, das zur Ermittlung der Gruppenzugehörigkeit stets nach dem Verfahren für den Mehr-Gruppen–Mehr-Variablen-Fall rechnet und die Klassifikation nach dem Wahrscheinlichkeitskonzept durchführt.

Dabei ergibt sich die Diskriminanzfunktion $y = 0{,}001\ x_1 + 0{,}004\ x_2$. Die Güte der Trennfunktion kann anhand des kanonischen Korrelationskoeffizienten bestimmt werden, der im vorliegenden Fall *0,670* beträgt. Eine andere

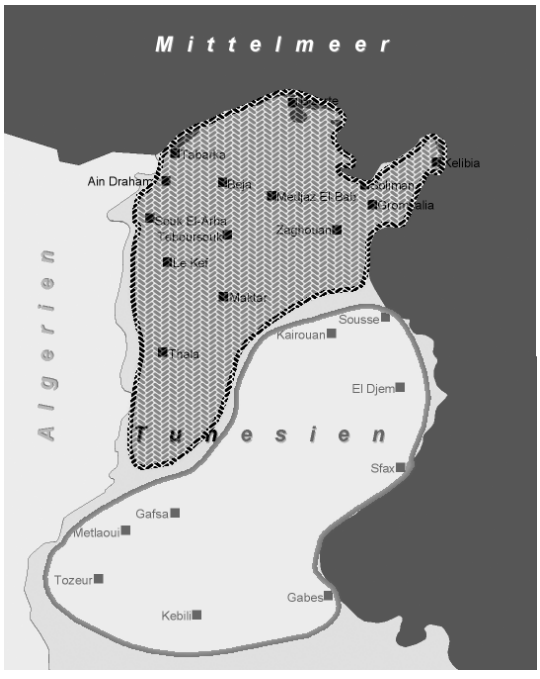

Abb. 5: Gruppierung der Klimastationen vor der Diskriminanzanalyse (Quelle: Eigener Entwurf)

Möglichkeit zur Feststellung der Trennkraft ist Wilks' Lambda, das in unserem Fall einen Wert von *0,552* hat.

Überprüft man nun die vorgegebene Gruppierung mit der errechneten Diskriminanzfunktion, so wird deutlich, das 2 Stationen falsch klassifiziert wurden: Die Stationen Medjaz El-Bab und Kelibia, die sich ursprünglich in der Gruppe A befanden, müssen nun in die Gruppe B einsortiert werden. Daraus ergibt sich nun folgende Klassifikationsmatrix:

		Gruppenzugehörigkeit nach der Diskriminanzanalyse			
		Cluster A		*Cluster B*	
Vorgegebene Gruppenzugehörigkeit	*Cluster A*	12	+	2	**= 14**
		+		+	
	Cluster B	0	+	9	**= 9**
		= 12		**= 11**	

Die Trefferquote, also der Prozentsatz der bereits richtig klassifizierten Stationen, beträgt in unserem Fall 91,3 %. Wie sich die Änderung der Klassifizierung geographisch auswirkt, zeigt **Abb. 6**.

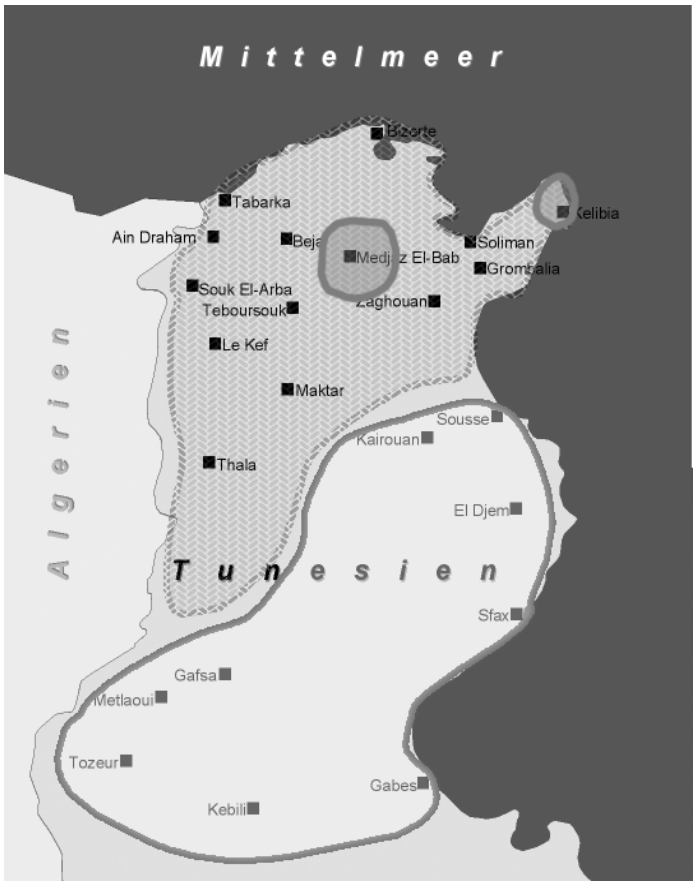

Abb. 6: Gruppierung der Klimastationen nach der Diskriminanzanalyse (Quelle: Eigener Entwurf)

4 Literaturverzeichnis

- Backhaus, K. et al.(2000[9]): Multivariate Analysemethoden. – Berlin, 90-163.
- Bahrenberg, G., E. Giese & J. Nipper (1992[2]): Statistische Methoden in der Geographie Band 2: Multivariate Statistik. – Stuttgart, 316 - 358.
- Rosner, H.-J. (1998[2]): Verarbeitung Geographischer Daten: Methodische Bausteine zu Statistik ung GIS. – Tübingen, 69 – 71.